ПРОГНОЗИРОВАНИЕ ПРОБОК НА УЛИЦАХ ПО ИЗВЕСТНЫМ ДАННЫМ О СКОРОСТИ АВТОМОБИЛЕЙ

Сергей Гуда, Денис Рябов
Южный федеральный университет, факультет математики, механики и компьютерных наук
e-mail: gudasergey@gmail.com, dryabov@yandex.ru

Аннотация

В рамках проводимого компанией «Яндекс» конкурса «Интернет-математика 2010» была предложена задача прогноза скорости движения автомобилей в Москве на основе имеющихся данных за один месяц. Для оценки качества предсказания использовалась определенным образом составленная невязка между известными и предсказанными данными, а победитель определялся как получивший минимальное значение невязки. В настоящей работе описывается алгоритм, который позволил сделать прогноз скоростей наиболее точно.

Ключевые слова: *прогнозирование пробок, Интернет-математика 2010.*

1. ВВЕДЕНИЕ

Рассматривается задача прогноза скорости движения автомобилей по улицам города, если известны только данные о скорости за предыдущий месяц и граф дорог. Данная задача была поставлена на конкурсе «Интернет математика 2010», проводимом компанией Яндекс (см. [1]). Количество машин или их плотность неизвестны. Данные о скорости автомобилей зачастую являются нерегулярными, на большинстве улиц имеются пробелы в данных величиной более получаса, а в некоторые дни данных нет вовсе. В некоторых случаях имеются противоречащие друг другу данные. Недостающую информацию можно почерпнуть с соседних улиц, благо граф дорог города учитывает правила дорожного движения (запрещенные повороты и проезды). Задача предсказания усложняется наличием некоторой хаотичности в данных, а также, по всей видимости, ошибок связанных с определением улицы по полученным от автомобильных GPS-навигаторов координатам.

Для оценки качества предсказания используется норма разности предсказанного и известного значения скорости в пространстве ℓ_1 с весовыми коэффициентами, которые делают предсказания скорости на длинных улицах и на больших временах более «ценными», чем все остальные. Такая оценка определяет используемые для предсказания методы.

Хорошо зарекомендовавшие себя методы предсказания пробок, основанные на анализе статистики плотности и маршрутов автомобилей (см. [2–3]) неприменимы к данной задаче ввиду отсутствия указанных данных. Поэтому единственное, что остается — строить предположение о скорости в заданный день на заданной улице исходя из данных в предыдущие дни на этой и на близлежащих улицах. Чтобы нейтрализовать стохастику в данных, для предсказания в момент времени t выбиралась окрестность $(t-r,t+r)$, и учитывались данные во все дни из заданной окрестности. Благодаря особенностям выбранной для оценки предсказания нормы лучшим является не среднее арифметическое значения скорости, а среднее медианное.

Чтобы учитывать данные различных дней с разными весами (например, будней и выходных), было решено перейти к вероятностным распределениям скорости. Выбор важности (веса) предыдущих дней для предсказания на требуемый день (здесь и далее — день X) является отдельной задачей. Ясно, что если день X — будний (а предварительный анализ данных показывает, что день X — понедельник), то выходные нужно брать с меньшим весом, чем будние дни. Также кажется логичным с большим весом брать дни того же дня недели, что и день X. Обоснование метода и выбор веса дней описаны в п. 4.1.

Помимо данных о скорости в предыдущие дни, в задаче также известна скорость в течение двух часов в день X. Это позволяет существенно улучшить качество предсказания. Об этом рассказано в п. 4.2.

Разные улицы обладают разными свойствами, а значит, для каждой нужны свои методы предсказания или свои параметры метода предсказания. К сожалению, имеющихся данных о скорости за 1 месяц оказывается недостаточно для эффективной тренировки метода на каждой отдельно взятой улице, поэтому улицы объединялись в группы и подбирались свои параметры для каждой группы. Использовалось несколько различных типов разбиений улиц на группы: по пропускной способности, по длине, по средней скорости, по количеству данных, по вкладу в оценку предсказания и т.п. Данной задаче посвящен п. 4.4.

В п. 5 рассказывается о свойствах и учете ошибок в имеющихся данных. Основная проблема состоит в копировании данных, полученных от GPS-навигаторов автомобилей на несколько различных улиц. Этот факт понятен для развязок, где несколько участков улиц расположены очень близко друг к другу. Однако, такое копирование наблюдается и для отдаленных улиц. Это особенно четко выражено в

дни с небольшим трафиком. К сожалению, это приводит к тому, что для определенных классов улиц оценка предсказания улучшается, если предсказывать ошибки, т.е. предсказывать что данные будут скопированы на все улицы класса.

2. ПОСТАНОВКА ЗАДАЧИ

Заданы: ориентированный граф дорог, учитывающий правила дорожного движения, данные о скорости автомобилей с 16:00 по 22:00 за 30 последовательных дней, данные о скорости с 16:00 по 18:00 в следующий, тридцать первый день (день X), информация о длине участков улиц и об их пропускной способности. Требуется предсказать скорость в заданные моменты времени с 18:00 до 22:00 в день X так чтобы оценка

$$Q(v) = \frac{1}{n}\sum k_i k_l \left| v^* - v \right| \qquad (1)$$

принимала как можно меньшее значение. Здесь n — общее количество предсказаний; k_l — длина улицы, отнесенная к 120 м; k_l=1+τ/40 — временной коэффициент; τ — количество минут, считая от 18:00, v^* — наблюдаемая скорость; v — предсказанная скорость. Таким образом, в данной формуле учитывается, что на более поздних временах может отсутствовать корреляция с имеющимися на интервал 16:00–18:00 данными, и предпочтение отдается алгоритмам, позволяющим делать долгосрочные прогнозы о развитии ситуации на дорогах города.

Заданные значения скоростей — целые. В предлагаемом алгоритме предсказываемые значения также являются целыми, хотя на этапе тренировки параметров метода предсказания использовались нецелые предсказания с целью сделать функцию $Q(v)$ непрерывной.

Известно, что требуется предсказать скорость для улиц Москвы, но названия улиц неизвестны. Вместо названий даны идентификаторы (целые числа). При разработке алгоритма не проводилась попытка сопоставлять дуги графа с реальными улицами, чтобы метод легко можно было перенести и на другой город. Хотя знание реальных улиц могло бы улучшить метод предсказания.

3. ОСОБЕННОСТИ ИСХОДНЫХ ДАННЫХ

Данные о скорости в «нормальном» случае идут с интервалом в 4 минуты, но это наблюдается редко. Иногда интервал составляет 2 минуты и довольно часто он больше 4 мин. Легко определяются будние дни и выходные (см. рис. 1, 2). Здесь и далее на всех графиках скорость измеряется в км/ч, время — число минут от 16:00.

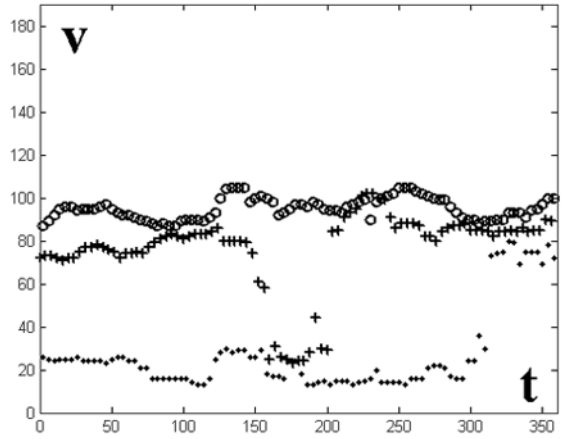

Рис. 1. Понедельник (крестики), пятница (кружки) последней известной недели на улице 843572

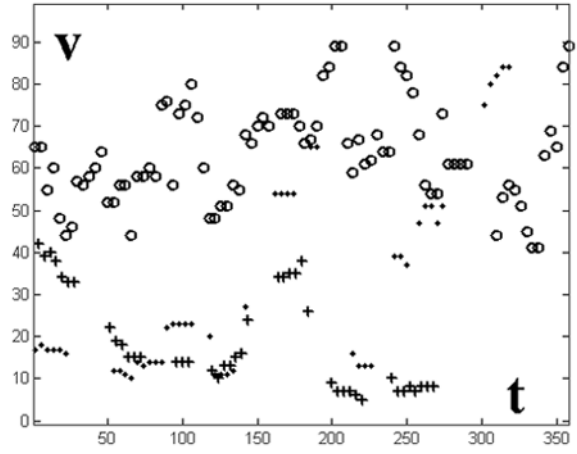

Рис. 2. Понедельник (крестики), пятница (точки) последней известной недели на улице 369115

Улиц с неполными данными много. Если взять за «нормальное» количество данных — число данных с интервалом в 4 мин., то для «нормальной» улицы за 30 дней должно быть задано 90*30=2700 измерений скорости. Процент улиц с числом данных менее 50% от «нормального» равен 82%. Чтобы оценить вклад этих улиц в общую оценку предсказания (1), воспользуемся определением.

Относительным вкладом множества улиц в оценку (1) назовем отношение

$$\frac{\sum_{\Omega} k_t k_l}{\sum k_t k_l}.$$

Знаменатель данной формулы является оценкой $Q(v)$ умноженной на n для предсказания v, отличающегося от наблюдаемого на 1; числитель — то же самое, только сумма берется по улицам из множества Ω. Относительный вклад улиц с малым числом данных в два раза меньше их количества — 40%. Это объясняется тем, что данных на таких улицах меньше, значит для них, скорее всего, известно меньшее число наблюдаемых значений скорости v^* в формуле (1).

Данные на некоторых улицах имеют хаотический характер (см. рис. 3). Причины такого поведения измерений неизвестны.

Иногда на фоне скорости 50–60 км/ч появляются всплески, превышающие 150 км/ч, что свидетельствует о явном нарушении правил дорожного движения такими водителями или погрешности в методике определения скорости.

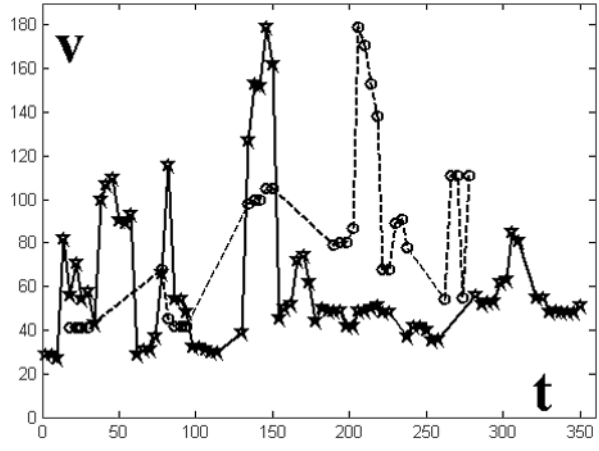

Рис. 3. Понедельник (сплошная) и воскресенье (пунктир) последней известной недели на улице 830658

Представляют интерес множества улиц с данными, совпадающими более чем на 50% в некоторые дни. Если предположить, что скорость может принимать с одинаковой вероятностью всего 10 различных значений, то для всех имеющихся данных вероятность того, что

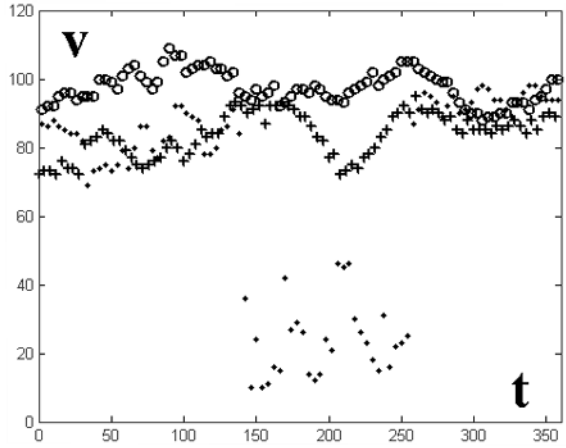

Рис. 4. Понедельник (сплошная), пятница (точки) и воскресенье (кружки) последней известной недели на улице 925236

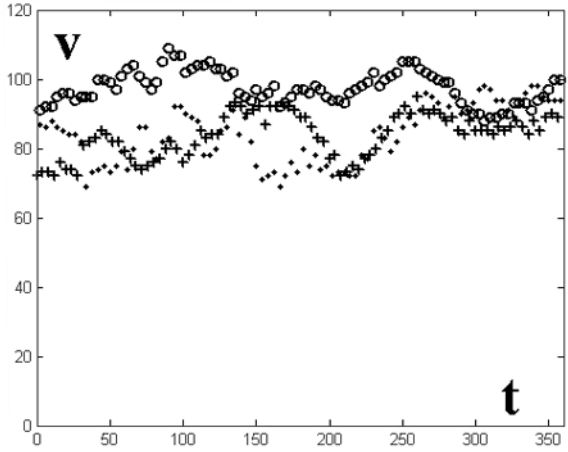

Рис. 5. Понедельник (сплошная), пятница (точки) и воскресенье (кружки) последней известной недели на улице 749501

два блока по 30 значений скорости совпадут, не более 10^{-18}. Это говорит о том, что найденные совпадения не случайны. Наличие улиц с совпадающими данными по-видимому связано с особенностями сбора информации о скорости автомобилей. Вероятно, для каждой улицы имеется окрестность около нее, поступающие данные из которой

осредняются по времени и пространству и заносятся в базу данных один раз в 4 минуты. Если автомобилей с GPS-навигаторами много, то у каждой улицы будут свои данные, а если мало, то у групп улиц данные будут совпадать. Если такое копирование скорости и вправду имеет место, то это является ошибкой в исходных данных задачи. На рисунках 4 и 5 изображены графики скоростей для двух разных улиц. Видно, что в воскресенье и понедельник они полностью совпадают, а в пятницу, когда на одной из улиц пробка, — нет.

4. АЛГОРИТМ ПОСТРОЕНИЯ ПРЕДСКАЗАНИЯ

4.1. Скользящее медианное среднее

Пусть нам требуется предсказать скорость в момент времени t^* дня X. Возьмем окрестность (t^*-r, t^*+r), где радиус r может варьироваться от улицы к улице (см. п.4.4). Рассмотрим значения скорости ai во все предыдущие дни в этот промежуток времени. Будем считать, что искомая скорость v^* — случайная величина, которая может принимать эти значения с вероятностями p_i. Тогда минимизация оценки (1) эквивалентна минимизации функции

$$q(v) = M(|v^* - v|)$$

где M — математическое ожидание. Данная функция достигает минимума в медианном среднем случайной величины v^*, т.е. такой точке m, для которой $P(v^* \le m) = P(v^* > m)$, где P — вероятность события.

Опишем выбор вероятностей $p_i = P(v^* = a_i)$. Зафиксируем один из 30 известных дней. Пусть $v(t)$ — скорость в этот день. Чтобы учесть возможность отсутствия данных на промежутке (t^*-r, t^*+r), использовался следующий подход. Возьмем функцию

$$\omega(t, t^*) = \frac{1}{ch^{\gamma}\left(\frac{t - t^*}{r}\right)}$$

с некоторой константой $\gamma > 0$. Положим

$$\tilde{p}_i = \sum_{t: v(t) = a_i} \omega(t, t^*)$$

Нормируя величины $\tilde{p}_i$, получим искомые вероятности для зафиксированного дня

$$p_i = \frac{\tilde{p}_i}{\sum_i \tilde{p}_i}$$

Теперь стоит задача о комбинации вероятностных распределений скорости в различные дни, для чего используется простейшая сумма

с некоторыми весами W, означающими важность дня в общей массе данных. Чтобы учесть зависимость веса от количества данных, попадающих в окрестность точки t^*, вместо вероятностей использовались исходные величины $\tilde{p}_i$. Если день X — будний, то ясно, что гораздо больший вес должны иметь будние дни, причем среди всех таких дней должны выделяться дни того же дня недели, что и день X. Существенное улучшение оценки дает повышение важности дня, предыдущего перед днем X, но это, однако, неприменимо к понедельникам.

4.2 Учет данных дня X

По условию задачи известны данные о скорости автомобилей с 16:00 до 18:00 в день X. Это позволяет существенно улучшить оценку (1).

4.2.1. Похожесть

По известной скорости в день X из всех предыдущих дней можно выбрать наиболее похожие и увеличить их вес. Автоматический подбор параметров, минимизирующих оценку (1), показал, что лучшим способом расчета весов W является вычисление нормы разности скорости v^x в день X и скорости v в известный день, а затем применение к этой норме функции φ, которая подбиралась автоматически. В итоге получим формулу

$$W = \varphi\left(\left\| v - v^x \right\|_{\ell_p}\right) \tag{2}$$

где

$$\left\| y \right\|_{\ell_p} = \left(\sum_i y_i^p\right)^{\frac{1}{p}}$$

функция φ для одной из улиц изображена на рис. 6.

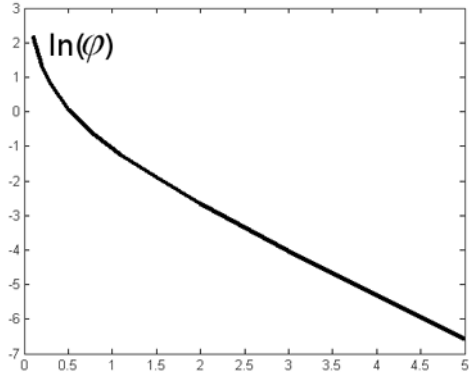

Рис. 6. Функция $\ln(\varphi)$

Значения скоростей v^x и v вообще говоря заданы в различные моменты времени. Для поиска недостающих значений применялась интерполяция. Лучшее значение индекса нормы p оказалось равным 50. Вероятно, если предварительно отчистить данные от больших редких всплесков скорости, лучшей будет норма в ℓ_∞: $||y||_{\ell_\infty} = \max|y_i|$.

Улучшение оценки (1) дает ввод временного веса при расчете нормы в (2). Похожесть данных в моменты времени ближе в 18:00 имеет большую важность, чем в более ранние (как видно из исходных данных, пробки на дорогах начинают возникать в интервал времени 17:00–18:30, при этом наличие данных в интервале 16:00–17:00 как правило слабо коррелирует с дальнейшим развитием ситуации на дорогах).

4.2.2. Экстраполяция данных дня X

Судя по графикам на рис. 1, 2 зависимость скорости от времени не является гладкой и имеет разрывы. Поэтому использовалась простейшая экстраполяция константой — средним нескольких последних значений скорости. Лучшее количество таких значений оказалось равным 2–3. Такую экстраполяцию лучше всего проводить только в случае, когда на улице началась пробка, и прогнозировать пробку необходимо в течение всего часа пик с 18:00 до 20:00.

4.3. Другие тонкости метода

Так как использованный подход в основе своей является статистическим, то улучшить его результат в первую очередь можно за счет более корректного задания определенных весов каждому из предыдущих дней недели. Несмотря на то, что основу рассчитанных весов составляет учет «похожести», в алгоритм были включены также следующие множители веса, которые также независимо подбирались для каждой группы улиц. Во-первых, был использован свой множитель веса для каждого дня недели. В частности, субботы и воскресенья брались с незначительным весом, достаточном лишь для того, чтобы улучшить ситуацию на тех улицах, для которых практически нет данных на будние дни недели. С понижающим множителем бралась пятница, так как по пятницам пробки длятся обычно дольше, чем в остальные дни. Во-вторых, для первых недель наблюдения были выбраны незначительные понижающие множители, и тем самым учитывался эффект «устаревания» имеющихся данных. Также для каждого дня недели незначительно варьировалась величина радиуса r. И наконец, был разработан метод, позволяющий для понедельников оценить наличие пробок в предыдущее воскресенье и на основе этого строить более точные предсказания.

4.3. Зависимость параметров метода от характеристик улицы и от времени

Введение своего набора параметров метода предсказания для каждой улицы позволяет существенно улучшить оценку. Таким образом учитывается зависимость параметров от некоторых характеристик улиц, их индивидуальных особенностей. Оптимальные значения параметров тренируются по известным данным о скорости за 1 месяц. К сожаленью, этих данных оказалось недостаточно для эффективной тренировки параметров метода. Поэтому пришлось объединять улицы в группы, и тренировать параметры для каждой группы. В итоге тренировка значений для половины из имеющейся сотни параметров метода привела к улучшению оценки предсказания.

Анализируя оптимальные параметры для групп, можно установить их зависимость от величин, используемых при разбиении улиц на группы. Так авторы выявили зависимость радиуса r окна осреднения (см. п. 4.1) от количества известных данных. Оказалось, что для улиц с большим числом данных уменьшение радиуса приводит к существенному улучшению оценки метода. Из общей теории (см. [4]) следует, что зависимость должна иметь вид $r \sim N^{-1/5}$, но предложенный нами подход является более общим и позволяет учитывать ситуации, когда могут иметь место отклонения от общей теории.

Увеличить точность предсказания также помогает введение зависимости параметров метода от времени, на которое делается предсказание. Была использована кусочно-постоянная зависимость с двумя или тремя значениями параметров для разных интервалов времени предсказания. Например, для улиц с большим числом данных оптимальный радиус r окна осреднения принимает три различных значения в зависимости от времени предсказания: с 18:00 до 20:00 r=3; с 20:00 до 21:00 r=4 и с 21:00 до 22:00 r=2. Оптимальный вес пятниц при оказался равным нулю.

5. СОВПАДЕНИЕ ДАННЫХ У РАЗНЫХ УЛИЦ

Во множестве известных данных есть улицы с частично совпадающими данными. Их больше всего в выходные, а также в первую половину известного месяца. Это объясняется естественными причинами — копированием данных на смежные улицы, если данных, поступающих от автомобильных GPS-навигаторов, мало.

Если искать улицы с данными, совпадающими только в воскресенье последней известной недели более чем на 50%, то получатся со-

вокупность множеств по 10–600 штук «похожих» улиц (улицы с числом данных меньше 20% от нормального в это воскресенье не учитываются). Для будних дней количество улиц с частично совпадающими данными гораздо меньше. Множества улиц с совпадающими данными используются следующим образом.

5.1. Способ 1: копирование предсказания на несколько улиц

Если вероятность копирования данных выше 0.5, то оценка улучшиться, если предсказывать это копирование. Весь временной интервал с 16:00 до 22:00 разбивается на две части: [17:45, 20:15] и [16:00, 17:45]∪[20:15, 22:00]. Для каждого интервала строятся множества совпадающих на нем улиц более чем на p процентов (p>50%). Во множествах улиц ищутся наиболее вероятные данные для каждого дня месяца (например, в каждый момент времени берется медианное среднее всех скоростей). Назовем их «центральными». Исходя из предположения, что это именно те данные, с которых происходит копирование на другие улицы, предсказание строится по этим данным, а затем для интервала времени множества улиц копируется на все улицы множества. Таким образом происходит удаление маловероятных собственных данных улиц, «мешающих» предсказанию.

Несмотря на то, что такое копирование предсказания приводит к улучшению оценки, это не является правильным. Но исправить положение может только изменение методики сбора информации о скорости автомобилей.

В качестве центральных данных можно использовать данные одной из улиц множества. Если ввести метрику ρ на множестве улиц как процент различных данных, то в качестве центральной логично взять улицу, между которой и большинством других расстояние ρ будет наименьшим. А именно, улица считалась центральной, если среднее медианное расстояний от нее до остальных улиц минимально.

Центральную улицу можно определить и другим способом. Центральной можно считать такую улицу, при которой описанная схема копирования предсказаний дает наилучший результат. С таким определение нужно быть осторожным. Оно заманчиво, но для выбора центра пробного предсказания на один известный день недостаточно. Нужно использовать предсказание на известную неделю.

5.2. Способ 2: оценка загруженности дорог

Совпадение с центральными данными в 41 день показывает, что на дороге мало машин и данные о скорости копируются, т.е. вероятность пробки невелика. Если же данные отличаются от центральных, то на улице точно будет пробка. Для использования такого подхода нужно аккуратно подбирать множества похожих улиц. Если множество окажется слишком широким, то оно может содержать несколько выраженных центральных данных. И выбор какого-то одного из них приведет к ухудшению предсказания. Если же множества окажутся слишком узкими, то улучшение оценки (1) будет незаметно. Предлагается строить множества улиц совпадающих более чем на 90% по субботе и воскресенью последней недели месяца.

6. БЛАГОДАРНОСТИ

Авторы выражают искреннюю благодарность Семену Кузьменко и Владимиру Ремизову за техническую поддержку проекта и удовлетворение все время возраставших потребностей в вычислительных мощностях.

ЛИТЕРАТУРА

1. Конкурс «Интернет математика 2010» http://imat2010.yandex.ru.
2. **J. Markoff**. Microsoft Introduces Tool for Avoiding Traffic Jams // New York Times, April, 2008.
3. **H. Kriegel, M. Renz, M. Schubert, and A. Zuefle**. Statistical Density Prediction in Traffic Networks // Proceedings of the 2008 SIAM International Conference on Data Mining, p.692.
4. **Хардле В.** Прикладная непараметрическая регрессия: Пер. с англ. - М.: Мир, 1993. - 349 с.
5. **R.Chrobok, A.Pottmeier, S.F.Hafstein, M.Schreckenberg**. Traffic forecast in large scale freeway networks // International Journal of Bifurcation and Chaos. Vol.14, No.6 (2004) 1995-2004.